Using the Institutional Repository to publish research data

Christopher Gutteridge

University of Southampton cjg@ecs.soton.ac.uk,
http://users.ecs.soton.ac.uk/cjg/

Abstract. For open research data to be fully utilised it must be discoverable. Many types of research dataset are impossible to identify by looking at them so metadata is essential. This is the only major issue with using existing Institutional Repositories to preserve and disseminate data. This paper suggests a simple scheme for facilitating discovery and reuse of open scientific data.

1 Introduction

> "The greatest crisis facing us is not Russia, not the Atom Bomb, not corruption in government, not encroaching hunger, nor the morals of the young. It is a crisis in the organization and accessibility of human knowledge. We own an enormous "encyclopedia" - which isn't even arranged alphabetically. Our "file cards" are spilled on the floor, nor were they ever in order. The answers we want may be buried somewhere in the heap, but it might take a lifetime to locate two already known facts, place them side by side and derive a third fact, the one we urgently need." - *Robert Heinlein, 1950*

It is already possible to use an Institutional Repository (IR) to store and disseminate datasets, but not yet common practice. This data will be most valuable when similar datasets from around the world are aggregated and all scientists can have access to all available data.

Best practice will require Open formats, Open licenses, provenance and discoverability. Google, and other search engines, may have solved many of the problems of finding text, but raw data is a stickier challenge, as it may be nothing more than a grid of numbers or other arcane formats. This paper suggests a simple mechanism to make all open datasets discoverable and recommends that data is stored in Institutional Repositories to provide reliable long term curation and availability.

2 Background

Most data created in research is not yet available online. The infrastruture to enable this already exists in the form of Institutional Repositories. Making raw

data available online allows it to be reused, and also supports the scientific process by allowing peers to repeat the analysis of the data and verify conclusions in a paper. It is impractical for an IR manager to do more than curate the data. Individual research communities will form their own practices, tagging data in their local IR to allow it to be discovered by other members of their community.

There is limited space in a print medium. This makes it impractical to include many pages of data which will be of interest to very few readers. Much research is created, reviewed and consumed without ever entering hard-copy, and digital media requires fewer physical restrictions.

15 years ago, Stevan Harnad made his Subversive Proposal[1] that the scholarly community should be sharing its research online, without barriers, plus a sketch of how to get there. This is now well under way and at the time of writing, 63% of publishers now permit some form of a paper to be made available online.[3]

There are many issues with the communication of research data including collection, provenance, curation, interoperation and dissemination.

3 Immediate Solution

There are nearly 1000 insitutional-style repositories in the world[2] and that number is increasing. Rather than design complex new systems, the remit of IRs should be expanded to include research data along with research outputs. Nothing is required but a change in repository policy, and the addition of a new option "dataset", plus encouraging researchers to deposit.

Many repositories already support a record type of "other" and using this is better than nothing, but the data will be difficult to discover.

4 Making datasets discoverable

A repository should allow the depositor to identify the content of a dataset. This needs to be painless, but must not be from a limited set curated by the library as that will be a barrier to the evolution of scientific communities. There is no reason that a repository cannot offer auto-completion of the field from a short list, so long as it does not preclude free text input.

The ideal solution would be to identify the subject and format of the contents of the dataset by one or more URIs, but URIs are cumbersome. As a more practical solution, dataset contents can be identified by either a URI or a short text string which will be treated as part of a data format namespace `http://ds.eprints.org/ns/`, giving a URI without requiring the depositing scientist to remember the whole thing. These identifiers can signify either the content of the dataset (e.g. a analysis of a crystal structure), the format (e.g. chemical .cif or .cml file), other properties such as strictness of protocols used, or any combination of these. Generally, an identifier will indicate both format and

content, but this will need to be established by various communities as their requirements will vary widely. Any dataset can be identified by multiple identifiers as appropriate.

These identifiers should be included in any electronic dissemination of the metadata of the repository. Including via the OAI-PMH protocol using `dc:subject`, in any RDF or RDFa relating to the record.

Table 1. Some possible identifiers, at different levels of granularity

`http://ds.eprints.org/ns/chem`	some kind of chemical data
`http://ds.eprints.org/ns/chem-cml`	a chemical in CML format
`http://ds.eprints.org/ns/chem-cml-cry`	a crystal in CML format
`http://ds.eprints.org/ns/chem-cml-cry-org`	an organic crystal in CML format

This system is deliberately a very loose semantic relationship. A single identifier can indicate any or all of type, subject and format. To keep the system managable for each community, as few identifiers as possible should be used. In the above example the 'chem-cml-cry-org' is almost certainly an unhelpful level of detail, and just 'chem' is no more useful than a Library of Congress subject. A good balance should be to use identifiers which indicate scientific area and the format of the data at a level of detail useful to aggregation tools. If the dataset is in a machine processable format, such as CML, then the identifier should reflect this to allow aggregation tools to discover and process these datasets. Later, if needed, semantic information can be returned from ds.eprints.org, naming established identifiers and indicating that they imply that a dataset has certain subjects, types and formats.

Ideally all data should be made available in Open formats, with Open licenses. Formats and licensing are essential for providing aggregation services and using the data in future work.

These changes can and should be made to the standard release of repository tools such as EPrints, D-Space and Fedora.

4.1 Very Large Datasets

Some datasets may be much too large to make available via HTTP, due to both expense of bandwidth, and it being impractical to download. Anything larger than a few terabytes cannot yet be usefully made available via the web in raw form, but that threshold will increase in time. The existence of large datasets can still be described in the IR, even if the URI identifying the dataset is not resolvable. In this case the record should also contain human-readable information on how to gain access to the dataset.

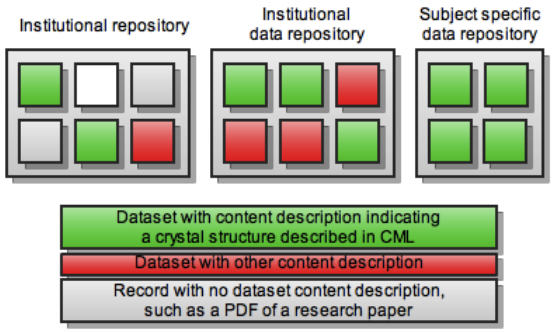

Fig. 1. Identifying datasets in various types of repository. Crystalography used as an example.

4.2 Datasets Requiring Additional Metadata

If the raw dataset just does not contain enough information to be useful, then the communities need to establish better formats. An interim solution is to make the URL of the dataset return a manifest file in XML, RDF or similar which contains the additional data to make the dataset useful.

4.3 Datasets with Multiple Files

Some datasets may contain multiple files. In this case, as with metadata being required, a manifest file can contain the URLs of the other files in the dataset. If the files have fixed names, then a less robust solution would be to load the other files based on relative URL paths. OAI-ORE[6] is suitable for the purpose and already supported by some repositories.

5 Policy

While some researchers are keen to publish their research papers online, many require a mandate before they will do so as it is one extra task for busy people. To ensure the majority of scientific data is made available will require mandates from funding bodies, accreditation exercises or institutions. When they emerge they should require that the data must not only be Open, but discoverable. Without clues to the content, search tools and harvesters may not be able to tell what many datasets are about.

Initially much data will not be in Open formats, but communities will discover benefits in standardising formats, but only if there is a practical way to aggregate the data.

For the first few years, it is very likely that many researchers will resist putting data online as they do not feel it is of a standard to publish, although they based published papers on it. Once it becomes part of the expected process, then this issue will diminish. To ease the process, researchers should be given credit for the quality and impact of their raw data, as well as their papers. Research papers should reference datasets used, both in the text and in electronic metadata.

6 Aggregators

With appropriately discoverable datasets available, the next step will to build aggregators which can add value to specific types of dataset. It is impractical for an IR to provide subject-aware visualisations and search tools, as there will be many and they will evolve. A more practical approach is to make the datasets available and discoverable with licences which make it possible for subject-specific web sites to provide these services over all datasets of a given type, from all the IRs in the world.

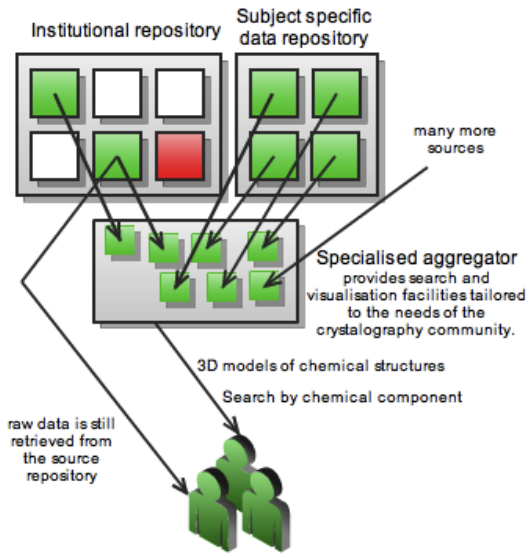

Fig. 2. An aggregator provides subject-specific value which which would be impractical for the generalist repositories which merely curate and disseminate the raw data.

7 Using the IR as storage for subject-specific or experimental services

While the Institutional Repository can provide a high degree of security in the continuity of URLs and preservation, there may be reasons for researchers to want to deposit their data in other systems. If these are well supported and stable then this is not a problem, however if these are more experimental then this is a concern. It is likely that many research projects will set up data repositories with no clear plan for continuity past the end of these projects.

A better solution, in many cases, is to use an institutional repository as a back-end to store the data. The experimental tool can both deposit items in the IR and then retrieve the data to provide subject-specific features or analysis. Where possible, it should disseminate the URI/URL of the raw data in the IR to proof against the risk of the experimental service going offline once funding ends.

A subject specific tool may act as both a tool for ingest (see fig. 3), and as a way to add value to specific types of dataset, in the same manner as an aggregation tool. The SWORD protocol[4] is ideal for this purpose. Such subject specific tools may well be commercial, or even supplied as an integrated solution with the next generation of laboratory devices.

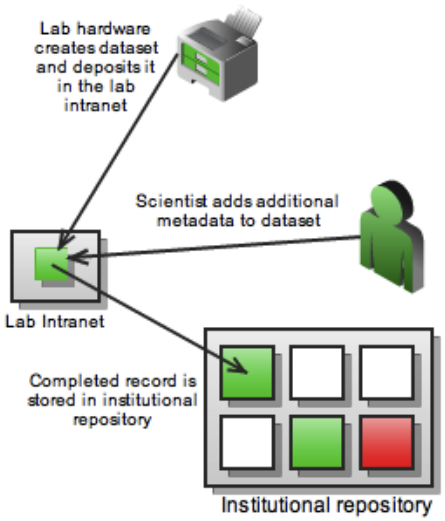

Fig. 3. Using a subject-specific tool to facilitate creation and storage of a dataset.

8 Alternatives to URI tagging raw datasets

The advantage of the author tagging the dataset is that it gives a degree of trust, as the information is collected and disseminated by an IR and that means the institution has a vested interest in ensuring that its data is correct and as-described. A downside is that if the community evolves new ways to identify its data, it is unlikely that anyone will update these tags.

An alternative or complimentary solution would be to use a social tagging system, such as Delicious[5] to tag the URLs or even URIs of datasets, either by a community effort (crowd-sourcing) or by a small expert group. However crowd-sourced information may be less reliable and in small communities is liable to noise. On the other hand, a single authority creating a global list of datasets for a given format is a single point of failure.

Social Bookmarking or authorities maintaining lists of known datasets are complementary to using subject tags or URIs to identify datasets. It is certain that different solutions will work for different communities.

9 Conclusion

Research funders should mandate that raw data produced as a result of their funding should be made available in Open formats, with Open licenses and made suitably discoverable at URLs which will be stable for many years. For this to be possible the researchers must have a sutiable repository for their data, and research communities will need to decide what level of detail is useful to facilitate discovery.

References

1. Harnad, S.: A Subversive Proposal. Association of Research Libraries. (1995) http://www.arl.org/sc/subversive/
2. Registry of Open Access Repositories (ROAR). http://roar.eprints.org/
3. SHERPA/RoMEO - Publisher copyright policies & self-archiving. http://www.sherpa.ac.uk/romeo/statistics.php
4. SWORD protocol. http://www.swordapp.org/
5. Delicious - social bookmarking tool. http://delicious.com/
6. Open Archives Initiative Object Reuse and Exchange (OAI-ORE). http://www.openarchives.org/ore/